Research

Open Access

Origins of chromosomal rearrangement hotspots in the human genome: evidence from the *AZFa* deletion hotspots

Matthew E Hurles*†, David Willey†, Lucy Matthews† and Syed Sufyan Hussain†

Addresses: *Molecular Genetics Laboratory, McDonald Institute for Archaeological Research, University of Cambridge, Downing Street, Cambridge, CB2 3ER, UK. †The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK.

Correspondence: Matthew E Hurles. E-mail: meh@sanger.ac.uk

Published: 14 July 2004

Genome ***Biology*** 2004, **5:**R55

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/8/R55

Received: 23 April 2004
Revised: 2 June 2004
Accepted: 7 June 2004

Abstract

Background: The origins of the recombination hotspots that are a common feature of both allelic and non-allelic homologous recombination in the human genome are poorly understood. We have investigated, by comparative sequencing, the evolution of two hotspots of non-allelic homologous recombination on the Y chromosome that lie within paralogous sequences known to sponsor deletions resulting in male infertility.

Results: These recombination hotspots are characterized by signatures of concerted evolution, which indicate that gene conversion between paralogs has been predominant in shaping their recent evolution. By contrast, the paralogous sequences that surround the hotspots exhibit little evidence of gene conversion. A second feature of these rearrangement hotspots is the extreme interspecific sequence divergence (around 2.5%) that places them among the most divergent orthologous sequences between humans and chimpanzees.

Conclusions: Several hominid-specific gene conversion events have rendered these hotspots better substrates for chromosomal rearrangements in humans than in chimpanzees or gorillas. Monte Carlo simulations of sequence evolution suggest that extreme sequence divergence is a direct consequence of gene conversion between paralogs. We propose that the coincidence of signatures of concerted evolution and recurrent breakpoints of chromosomal rearrangement (mapped at the sequence level) may enable the identification of putative rearrangement hotspots from analysis of comparative sequences from great apes.

Background

The pattern of meiotic homologous recombination is not homogeneous throughout the human genome. Hotspots of recombination activity - short genomic regions defined at the sequence level that exhibit higher levels of recombination than their surrounding sequence - have been identified in both the generation of haplotypic diversity by allelic homologous recombination (AHR) and the production of chromosomal rearrangements by non-allelic homologous recombination (NAHR) [1,2]. The evolution and determinants of these recombination hotspots are poorly understood.

NAHR between duplicated sequences sponsors a wide variety of pathogenic chromosomal rearrangements, giving rise to

phenotypes known collectively as 'genomic disorders' (reviewed in [3]). While some of these disorders result from the intermediates of NAHR being resolved as crossovers (that is, rearrangements such as deletions, duplications and inversions), others arise from the same intermediates being resolved as gene conversion events. Gene conversion is the non-reciprocal exchange of sequence between homologous sequences. Genomic disorders can arise when a gene conversion introduces a deleterious mutation into a functional gene, often from a reservoir of mutations present within a pseudogene. The results of NAHR need not be pathogenic, but can result in structural polymorphism (for example, the Yp paracentric inversion [4]). Similarly, gene conversion is capable of homogenizing both allelic and non-allelic (paralogous) sequences with no deleterious consequences.

In contrast to AHR, where unknown factors maintain recombination hotspots at specific locations [5], the frequency of NAHR is thought to be related at least in part to the sequence identity between paralogous (duplicated) sequences [3]. Studies of pathogenic rearrangements [6] and cell-culture assays [7,8] have demonstrated the length of sequence identity to be a primary determinant of NAHR.

It has recently been suggested that gene conversion between paralogs may render them better substrates for NAHR as a result of increased sequence similarity [9]. Evidence for gene conversion between paralogs has been detected at a number of loci known to be involved in sponsoring pathogenic chromosomal rearrangements [9-12]. If these patterns of gene conversion are polymorphic, some individuals may manifest increased germline rates of specific chromosomal rearrangements. Similarly, over an evolutionary timescale, gene conversion events may result in the formation of species-specific rearrangement hotspots.

In evolutionary analyses, the imprint of gene conversion can be recognized through the distinctive pattern of 'concerted evolution' [13], whereby duplicated sequences within one species (paralogs) are more similar to one another than either is to their ortholog in a closely related species. In principle, two different mechanisms can cause concerted evolution: gene conversion or unequal crossing over within a tandem array of repeats.

To investigate the role of gene conversion in the evolution of NAHR hotspots it is necessary to carry out comparative sequencing of paralogs. The presence in four copies of even minimally duplicated paralogs on autosomes complicates comparative sequencing strategies, and so it was decided to focus on loci on the constitutively haploid portion of the Y chromosome.

NAHR between two paralogous HERV15 proviral sequences (of approximately 10 kb) flanking the Y-chromosomal *Azoospermia Factor a* (*AZFa*) locus produces deletions causing male infertility [9,14,15]. All sequenced *AZFa* rearrangement breakpoints in these sequences fall within two NAHR 'hotspots' - ID1 and ID2 - that are characterized by elevated levels of sequence similarity.

In this study, comparative sequences were obtained for the entire proximal and distal *AZFa*-HERVs from common chimpanzee (*Pan troglodytes*) and gorilla (*Gorilla gorilla*). The evolutionary history of sequences within the two rearrangement hotspots was compared to that of the remainder of *AZFa*-HERV sequences outside hotspots from the same species. Gene conversion simulations were developed to ask how well these empirical comparisons reflect the dynamics of the underlying gene conversion process.

Results and discussion

We sequenced the proximal (approximately 10 kb) and distal (approximately 12 kb) *AZFa*-HERVs from a male chimpanzee (GenBank accession numbers AY573558 and AY573559) and a male gorilla (AY573560 and AY573361). These sequences encompass two intervals containing the NAHR hotspots ID1 and ID2 [14], in which all 15 sequenced NAHR breakpoints have been found in humans (Figure 1, Table 1). ID1 and ID2

Table 1

The 15 characterized NAHR breakpoints within *AZFa*-HERVs

Study	Type of rearrangement	Independent rearrangements	Breakpoint in ID1	Breakpoint in ID2
[9]	Deletion	1	0	1
[9]	Gene conversion*	2	2	2
[14]	Deletion	6	4	2
[15]	Deletion	2	1	1
[25]	Duplication	2	2	0
		Total	9	6

*These gene conversion events appear to be associated with a double-crossover mechanism, and so each gene conversion event is associated with two breakpoints (see Figure 10b).

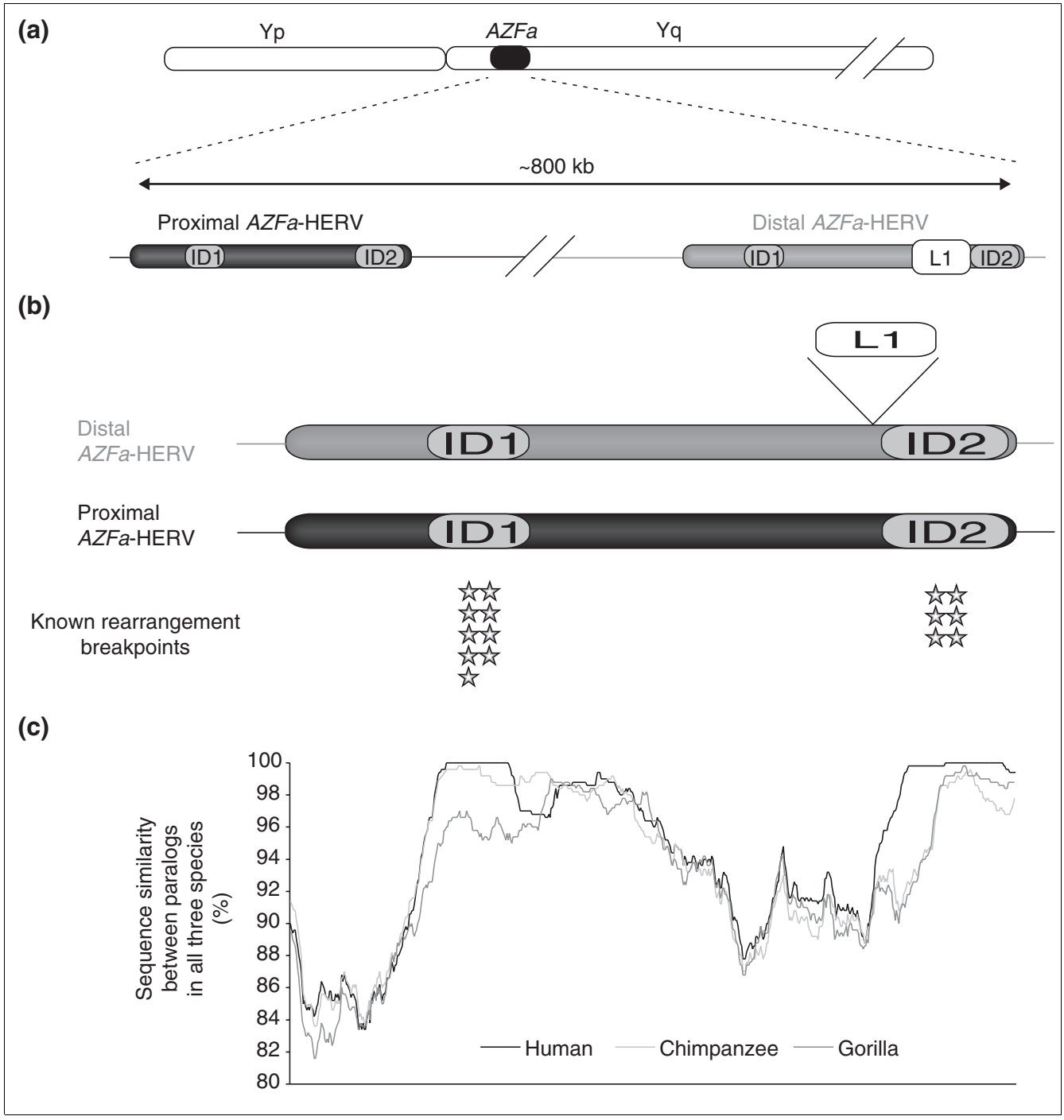

Figure 1
The *AZFa* locus and flanking *AZFa*-HERV proviral sequences. **(a)** The *AZFa* locus on Yq is flanked by two HERV15 proviral sequences in direct orientation. **(b)** The proximal and distal *AZFa*-HERVs are aligned with one another, with the inserted L1 material excised. The 15 known rearrangement breakpoints are shown by stars and fall in the ID1 and ID2 hotspot intervals within the *AZFa*-HERVs. **(c)** The plot of similarity between *AZFa*-HERV sequences in each of the three species shows that these hotspots are coincident, with greatest sequence similarity in humans, but that the sequence similarity in chimpanzees and gorillas is lower at these hotspots than in human.

comprise respectively 1,285 base pairs (bp) of absolute identity between paralogs and 1,690 bp containing a single 'paralogous sequence variant' (PSV). These are by far the longest stretches of sequence identity in the *AZFa*-HERVs, as revealed by the paralog similarity plot in Figure 1.

There are longer stretches of sequence identity in humans than in chimpanzee or gorilla

Strikingly, the longest stretches of identity between *AZFa*-HERV paralogs are much smaller in the chimpanzee and gorilla than in humans (Figure 1b, Table 2). This finding suggests that either gene conversion has homogenized the human *AZFa*-HERVs, or the human sequences have retained longer stretches of ancestral paralog identity that have diverged in chimpanzees and gorillas.

Table 2

Lengths of identity in the NAHR hotspots in different species

Species	Longest stretch of identity (bp) between paralogs in	
	ID1	ID2
Human	1285	1242
Chimpanzee	300	207
Gorilla	166	377

Detecting the signature of concerted evolution

We examined the expectation that in the presence of gene conversion a phylogeny comprising both paralogous and orthologous sequences should show clustering of paralogs, whereas in its absence, orthologs should cluster. If we construct a phylogeny from concatenated *AZFa*-HERV sequences excluding the two rearrangement hotspots ('Non-ID'), the orthologs cluster, indicating an absence of gene conversion (Figure 2a). This is possibly the result of paralog sequence similarity being below a required threshold for the initiation of homologous recombination. By contrast, all three paralogous sequences cluster in the phylogeny of the sequences of the ID1 hotspot. The ID2 phylogeny exhibits a third topology in which the human paralogs cluster, but the chimpanzee and gorilla paralogs do not.

An alternative phylogenetic method for exploring the impact of gene conversion is to examine the degree to which there are conflicting phylogenetic signals in the sequence data. If there has been only partial gene conversion throughout an alignment, then a subset of sites at which orthologous sequences are more similar than paralogous sequences will remain. If these sites predominate, a phylogenetic tree will show ortholog clustering, and the evidence of gene conversion will have been disregarded. Split decomposition disentangles these conflicting signals and allows the data to be displayed as phylogenetic networks. A tree-like network will result when no conflicting signals are apparent, but if conflicting signals from the common ancestry of orthologs and gene conversion between paralogs are present, four-sided cycles should appear in the network.

Figure 2b shows phylogenetic networks for the hotspot and non-hotspot portions of the *AZFa*-HERVs. The non-hotspot portion presents a tree-like network with only slight evidence of cycles, in which orthologs cluster together, suggesting that there has been minimal gene conversion in these intervals. In contrast, both ID1 and ID2 intervals present networks containing large cycles. In the case of the ID1 network, the presence of cycles despite the clustering of paralogs suggests that while gene conversion has not eradicated all evidence of the common ancestry of orthologs over the entire region, it nonetheless predominates in all three species. The phylogenetic network of ID2 has substantially larger cycles than that of ID1, which, together with the incomplete clustering of paralogs identified above, indicates that the conflicting signals of the common ancestry of orthologs and gene conversion between paralogs are more equally balanced in the ID2 interval. This suggests that gene conversion has been less pervasive at ID2 than at ID1. The extremely tight clustering of the human ID2 paralogs suggest that this paucity of gene conversion is confined to chimpanzees and gorillas, and more evidence to support this hypothesis is presented later.

Conflicting phylogenetic signals from individual phylogenetically informative sites are apparent in site-by-site compatibility analyses of both ID1 (data not shown) and ID2 (Figure 2c) alignments. The patterns of variant sites within these alignments suggest that the hotspot intervals have not been generated by single, long conversion events, but rather are patchworks of several shorter gene conversion events of opposing directions (proximal to distal, and distal to proximal).

The ID1 and ID2 hotspots were sequenced in another male chimpanzee and gorilla, and identical phylogenetic signals of concerted evolution were obtained. Thus we can infer that gene conversion since the human-chimp common ancestor has homogenized these *AZFa*-HERVs in humans, generating the recombinogenic intervals ID1 and ID2.

Greater sequence divergence between orthologs within NAHR hotspots

Comparisons of sequence divergence between the different orthologs in the three species reveal surprisingly high levels of sequence divergence at the ID1 and ID2 hotspots (Figure 3a) when compared to both the non-hotspot portions of the *AZFa*-HERV sequences (Non-ID), and previously published data on a single copy Y-chromosomal locus, *SMCY* [16,17]. For example, the Jukes-Cantor distance between human and chimpanzee orthologs is 2.5% averaged over the ID1 and ID2 hotspots, in comparison to 1.7% for the SMCY locus and 1.8%

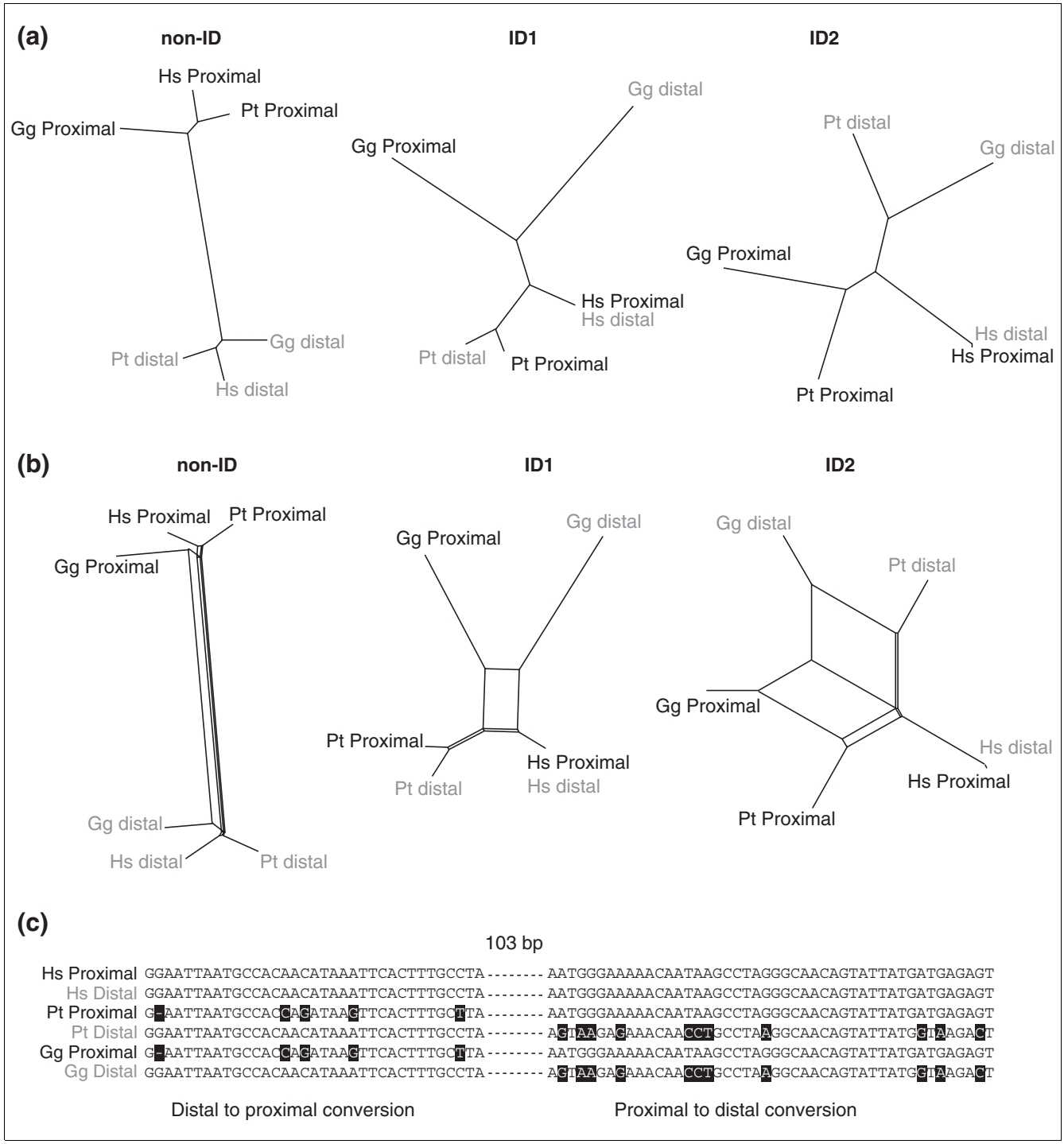

Figure 2
Evidence of concerted evolution at ID1 and ID2. **(a)** Neighbor-joining trees of comparative sequences over three intervals within the *AZFa*-HERVs (Pt, *Pan troglodytes*; Hs, *Homo sapiens*; Gg, *Gorilla gorilla*). **(b)** Phylogenetic networks of the same three intervals. **(c)** Two segments of an ID2 alignment demonstrating alternative directions of gene conversion events. Variant positions within the alignment relative to the human sequences are highlighted in black. In the first panel of the ID2 alignment, multiple variants specific to the proximal paralog in both chimpanzee and gorilla sequences are missing from humans, suggesting gene conversion on the hominid lineage using the distal paralog as a donor. By contrast, the second panel of the same alignment indicates gene conversion of the opposite directionality.

averaged over the Non-ID intervals. The average sequence divergence across these NAHR hotspots is greater for all three pairwise comparisons of hominoid species (human versus chimpanzee, human versus gorilla and gorilla versus chimpanzee) when compared to both the non-hotspot sequences and the SMCY locus. These latter two orthologous loci are indistinguishable, indicating that it is not paralogy *per se* that causes this elevated sequence divergence. The elevation in sequence divergence between the orthologous hotspot sequences is much less pronounced in comparisons between chimpanzee and gorilla. This difference can be attributed to a lack of elevated sequence divergence at ID2 between these two species (Figure 3b). The sequence divergence at ID2 between chimpanzee and gorilla is statistically indistinguishable to that between the single-copy SMCY loci in the same two species, and is substantially less than the divergence at ID2 between human and gorilla.

Rather than dividing up the *AZFa*-HERV alignment on the basis of prior information about the position of known rearrangement hotspots, it would be useful to detect regions of concerted evolution within an alignment of duplicated sequences in a *de novo* fashion. To this end, we have devised a statistic, the concerted index (CI), which varies between 0 and 1, revealing areas of low and high concerted evolution, respectively (see Materials and methods for details). This statistic can be calculated in sliding windows across an alignment. Applying this statistic to the *AZFa*-HERV sequence alignment reveals two zones of concerted evolution within the alignment, one containing ID1 and the other containing ID2 (Figure 4). These zones of concerted evolution contain the regions of the *AZFa*-HERV alignment that exhibit elevated sequence divergence between orthologs. The 5'-most zone extends some distance 3' to the ID1 hotspot. This observation accords with a recent finding that this region flanking the ID1 hotspot undergoes frequent gene conversion in humans [18], although no rearrangement breakpoints have yet been mapped to this conversion-only hotspot.

Simulations of the impact of gene conversion

The question arises of whether the unusual features of the observed data can be explained solely by an underlying gene conversion process. In addition, we are interested to explore how various parameters of gene conversion dynamics might affect sequence divergence among paralogs undergoing gene conversion. Consequently, Monte Carlo simulations were performed to explore the effect of gene conversion between paralogs on the evolution of duplicated sequences, according to a model shown schematically in Figure 5. Essentially, two 10-kb duplicated sequences have diverged by a specified amount before a speciation event. After speciation, the duplicated sequences continue to diverge, but random gene conversion events occur between the paralogs within each daughter species. These gene conversion events are restricted to the first half of the duplicated sequence, and conversion tract lengths are drawn from a geometric distribution, which accords well with empirical evidence [8,19,20]. The second half of the duplicated sequence, in which no gene conversion takes place, provides a control against which the impact of gene conversion can be measured.

Simulating the impact of varying gene conversion rates

Figure 6 shows the effect of varying the gene conversion rate on sequence similarity between paralogs and between orthologs. As expected, when the rate of gene conversion is zero, no difference can be observed in either paralog or ortholog similarity between the two halves of the 10-kb sequence. By contrast, the plot of paralog similarity shows that as the rate of gene conversion increases, the first halves of the paralogs become increasingly homogenized relative to the second halves (which are not exposed to gene conversion events). The plot of ortholog similarity shows that increasing the gene conversion rate increases the sequence divergence observed in the first half of the sequence over and above the sequence divergence observed in the portion of the sequence not exposed to gene conversion, despite the same rate of base substitution in the two halves. Other parameters within this model of gene conversion can be varied besides the gene conversion rate. Increasing the mean gene conversion tract length has a similar effect to raising the gene conversion rate (data not shown).

Our simulations indicate that the degree to which sequence divergence between orthologs is exaggerated by gene conversion between paralogs depends on the amount of gene conversion (as expressed by the rate and tract length of gene conversion events). In accordance with the results from our simulations, the zones of concerted evolution (high CI) contain the regions of the *AZFa*-HERVs displaying the highest ortholog sequence divergence. In general, it should be expected that the sequence divergence between human and gorilla orthologs should be similar to that between chimpanzee and gorilla orthologs at the same locus. However, this does not seem to be the case at ID2, where, as noted above, there appears to be less sequence divergence at ID2 between chimpanzee and gorilla than between human and gorilla. Thus, these differences in sequence divergence described above could indicate an increased rate of gene conversion at ID2 that is specific to the hominid lineage. No such effect is observable at ID1, where there is no discernible difference in sequence divergence between the human-gorilla and chimpanzee-gorilla comparisons. This inference of different gene conversion rates at ID2 between humans on the one hand, and chimpanzees and gorillas on the other, agrees with the analysis of phylogenetic networks above, where neither non-human species exhibited paralog clustering in phylogenies of ID2. Similarly, a plot of the CI in comparisons between chimpanzee and gorilla sequences shows a much narrower 3' zone of concerted evolution, again suggesting much less gene conversion in ID2 in these two species (data not shown).

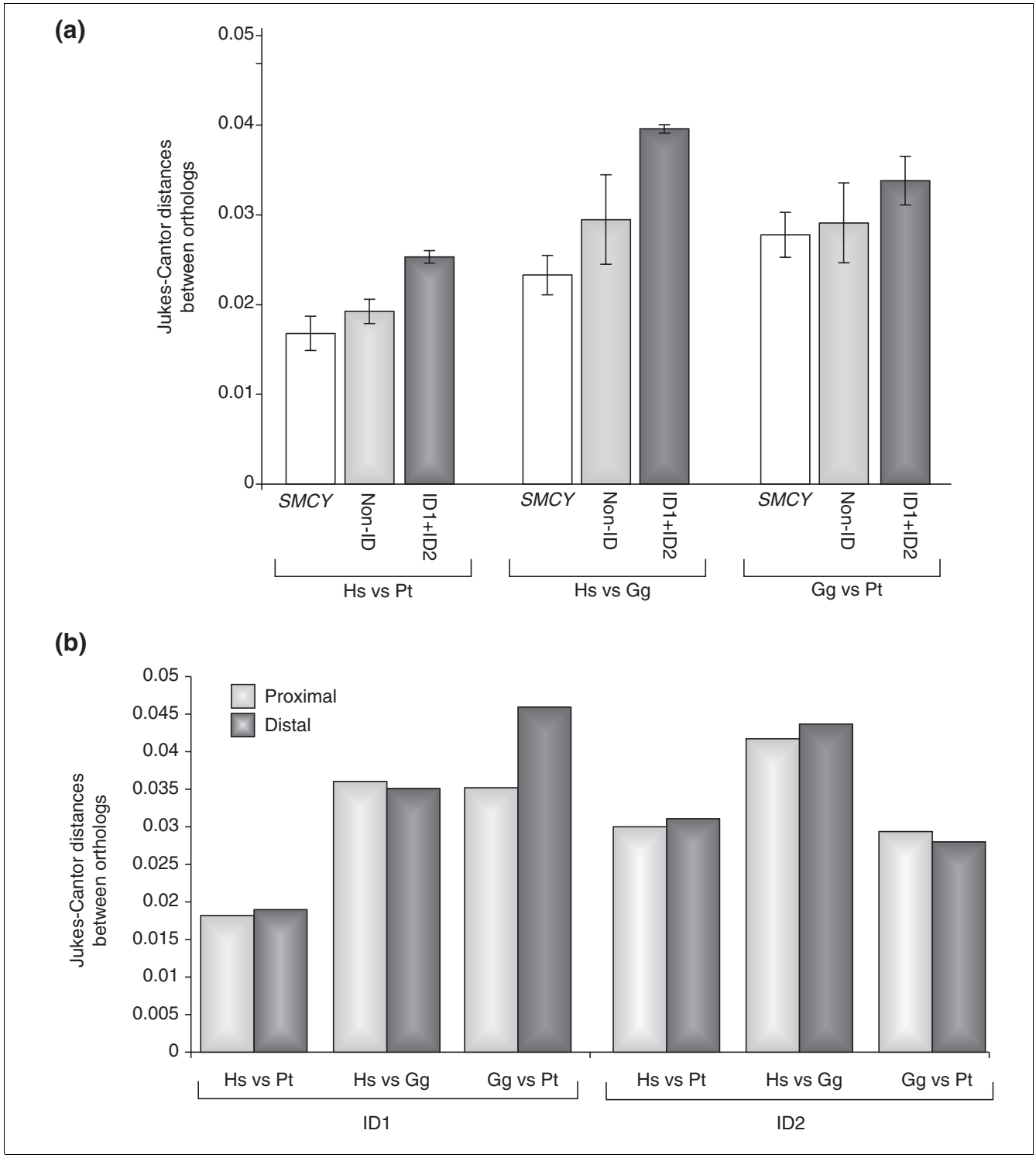

Figure 3
Sequence divergence between orthologous sequences within *AZFa*-HERVs. **(a)** Comparison of sequence divergence at three different sequence classes: two at *AZFa*-HERV (non-hotspot (non-ID) and hotspot (ID1+ID2)) and one from elsewhere on the Y chromosome (*SMCY*). Pairwise comparisons of three hominoid species (Pt, *Pan troglodytes*; Hs, *Homo sapiens*; Gg, *Gorilla gorilla*) were made. *AZFa*-HERV sequence divergences represent averages over both proximal and distal copies. The ID1 and ID2 sequence divergences are averaged to give ID1+ID2, rather than the sequence concatenated. **(b)** Sequence divergences between individual pairs of orthologous sequences at ID1 and ID2.

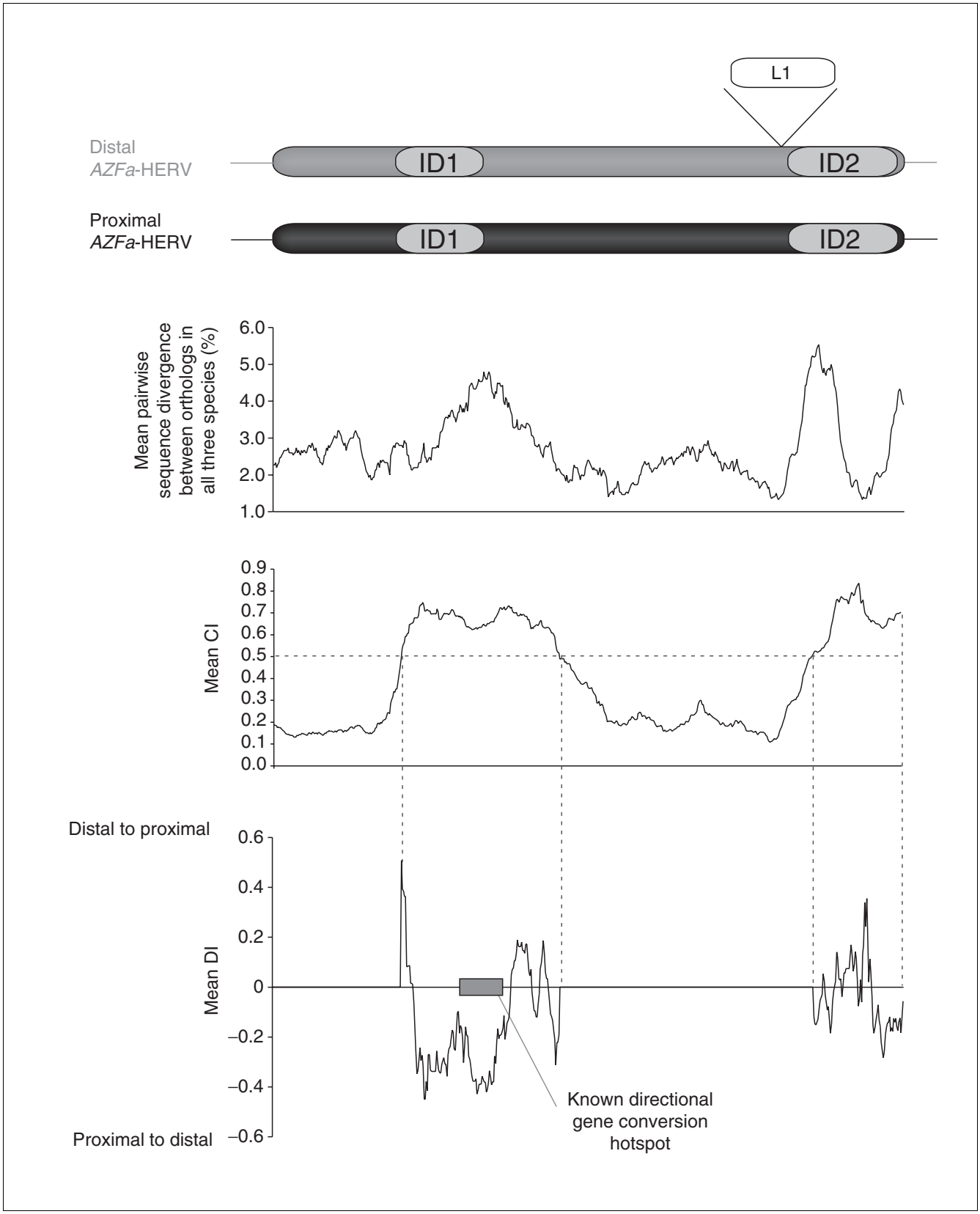

Figure 4 *(see legend on next page)*

Figure 4 *(see previous page)*
Sliding-window analyses across the *AZFa*-HERV alignment. Beneath the schematic alignment of the proximal and distal *AZFa*-HERVs (with the inserted L1 material excised) are sliding-window analyses showing how various sequence measures vary across the alignment. These statistics represent the mean of these values across all three pairwise comparisons. The three measures applied are: the orthologous sequence divergence, the concerted index (CI) and the directionality index (DI), where the CI is greater than 0.5 (see text for details).

Simulating the impact of varying initial paralog divergence

Running the simulations for different degrees of sequence divergence between the paralogs before the speciation event gives the results shown in Figure 7. A greater degree of sequence divergence between the paralogs before speciation leads to larger differentials in sequence divergence between the two halves of the paralogs. Similarly, the plot of ortholog similarity indicates that increasing the amount of variation between the paralogs before speciation leads to greater sequence divergence in the portion of the sequence undergoing gene conversion. The plot of ortholog similarity also shows that gene conversion only exaggerates sequence divergence between orthologs when the two paralogs have diverged before speciation. No increase in sequence divergence is observed in the portion exposed to gene conversion processes when the two paralogs are identical before speciation. This makes intuitive sense: if paralogous sequences are initially identical at speciation, gene conversion is dependent on the nucleotide-substitution process to generate new variants that can be gene converted from one paralog to another. However, the gene conversion process is equally capable of copying a novel derived allele from one paralog to the other as it is of reintroducing the ancestral allele from the other paralog. Therefore the rate of ortholog divergence does not differ from the nucleotide-substitution rate. However, if paralogs are not identical at speciation, each paralog contains a reservoir of variants that can be introduced into the other. Subsequent gene conversion events between the paralogs are more likely to introduce new sites that differ between orthologous sequences than they are to remove them. As a consequence, the occurrence of independent gene conversion events in both daughter species augments the normal nucleotide substitution process in generating sequence divergence.

Our simulations suggest that, given the same gene conversion rate, we should observe greater ortholog sequence divergence in portions of our *AZFa*-HERV alignments undergoing concerted evolution where paralog divergence at speciation was greatest. Clearly we are unable to sequence *AZFa*-HERVs from common ancestors of the species studied here. However, if we analyze portions of the *AZFa*-HERV alignment between human and chimpanzee that have a high CI (greater than 0.5) there does appear to be a strong correlation between mean ortholog divergence and mean paralog divergence (Figure 8). It is worth remembering that at some point paralog divergence at speciation will reach a threshold at which gene conversion is suppressed and no elevation in ortholog divergence would be observed.

Simulating the impact of varying gene conversion directionality

Further simulations were performed to explore the effect of introducing directionality into the gene conversion model, whereby gene conversion is not equally probable in both directions (proximal to distal, or distal to proximal). Figure 9 shows that although biased directionality has no effect on the degree to which paralogs are homogenized, increasingly biased directionality causes discrepant ortholog sequence divergence between proximal (p1 versus p2) and distal comparisons (d1 versus d2). Smaller perturbations of sequence divergence are observed in the preferred donor than in the preferred acceptor.

This observation from our simulations provides a means to investigate biased gene conversion directionality in our comparative sequence data by comparing proximal and distal orthologous divergences. In our *AZFa*-HERV alignments there are no consistent differences across all pairwise comparisons in proximal versus distal ortholog sequence divergences at either ID1 or ID2 (Figure 3b). This indicates that gene conversion at the *AZFa*-HERV hotspot intervals is relatively unbiased with respect to which paralog acts as the donor sequence. This finding is supported by the observation of bidirectional gene conversion events in the site-by-site compatibility analysis in ID2 described above (Figure 2c). We have developed a directionality index (DI) (see Materials and methods for details) to allow a sliding-window approach to detect zones of high directionality. A recent study has demonstrated the existence of a highly directional gene conversion hotspot that lies just 3' to ID1 in humans [18]. If we scan across the portions of the *AZFa*-HERV alignment with a high CI (greater than 0.5) and calculate this DI, we do indeed pick up an extended region of high proximal-to-distal directionality that corresponds to this conversion hotspot (Figure 4).

Conclusions

Association between an evolutionary history of gene conversion and chromosomal rearrangement hotspots

Several independent analyses of our comparative sequence data have revealed a clear history of concerted evolution within the *AZFa*-HERVs. Concerted evolution is mainly confined to the two rearrangement hotspot intervals within the *AZFa*-HERVs. The sequence identity that defines the ID1 and ID2 chromosomal rearrangement hotspots has arisen since the human-chimpanzee common ancestor as a result of several gene conversion events between proximal and distal *AZFa*-HERVs. This gene conversion process appears to be

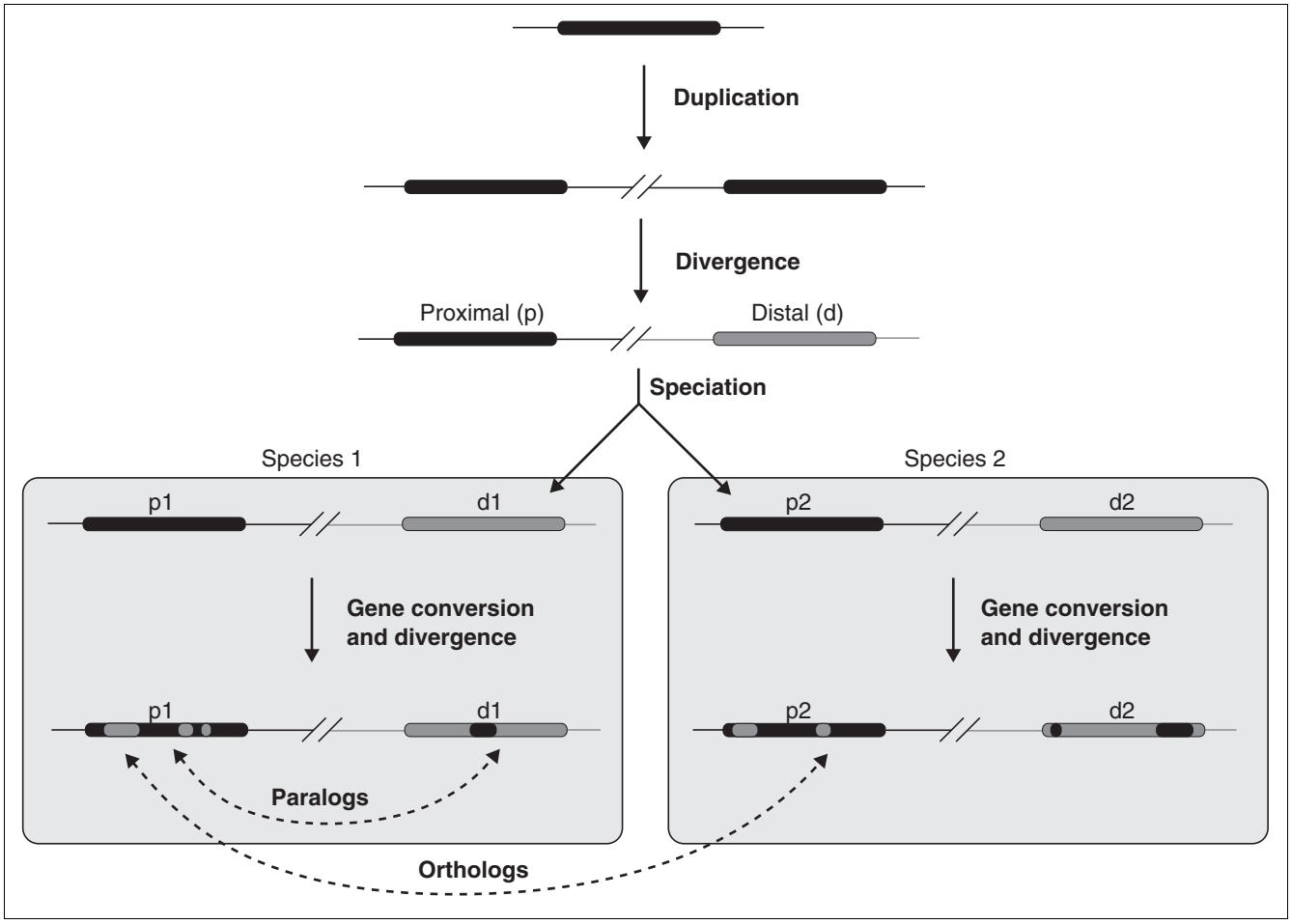

Figure 5
The model of sequence evolution incorporated into the simulations. See text for explanation. p1, proximal repeat in species 1; p2, proximal repeat in species 2; d1, distal repeat in species 1; d2, distal repeat in species 2.

unbiased with respect to which paralog acts as a donor sequence. In addition, there is good evidence, for one of these hotspots, of an acceleration of gene conversion in the hominid lineage.

It is tempting to ascribe a causal relationship to this association between chromosomal rearrangement hotspots and an evolutionary history of gene conversion. Under such a scenario, a rare gene conversion event homogenizes a small region of the *AZFa*-HERVs that hence becomes a better substrate for NAHR, which leads in turn to an increased frequency of further gene conversion events and chromosomal rearrangements. However, because gene conversion and chromosomal rearrangement reflect the alternative products of a common intermediate, it may be that a recombinogenic sequence motif underpins the association, and increased sequence identity has only a minor role in determining the frequency of chromosomal rearrangement. A lack of association between lengths of identity and rearrangement hotspots may also occur in recently duplicated sequences that have yet to diverge sufficiently in their non-recombinogenic portions for the hotspot-associated intervals of exaggerated paralog similarity to become apparent. An association between chromosomal rearrangement hotspots and an evolutionary history of gene conversion could also break down if there are sequence-dependent factors at NAHR hotspots which moderate the resolution of recombination intermediates such that crossovers are preferred to gene conversion events, leading to a higher frequency of rearrangement breakpoints within the hotspot, but not increased gene conversion. Identifying substantial numbers of *de novo* rearrangements and gene conversion events in pools of sperm will allow many of these issues to be resolved.

While our data are important for understanding the longer-term evolution of rearrangement hotspots, they are not ideal for estimating the parameters of inter-paralog gene conversion processes (for example tract length and rate) in the human genome. This is especially true when, as is the case here, gene conversion does not appear to have been constant

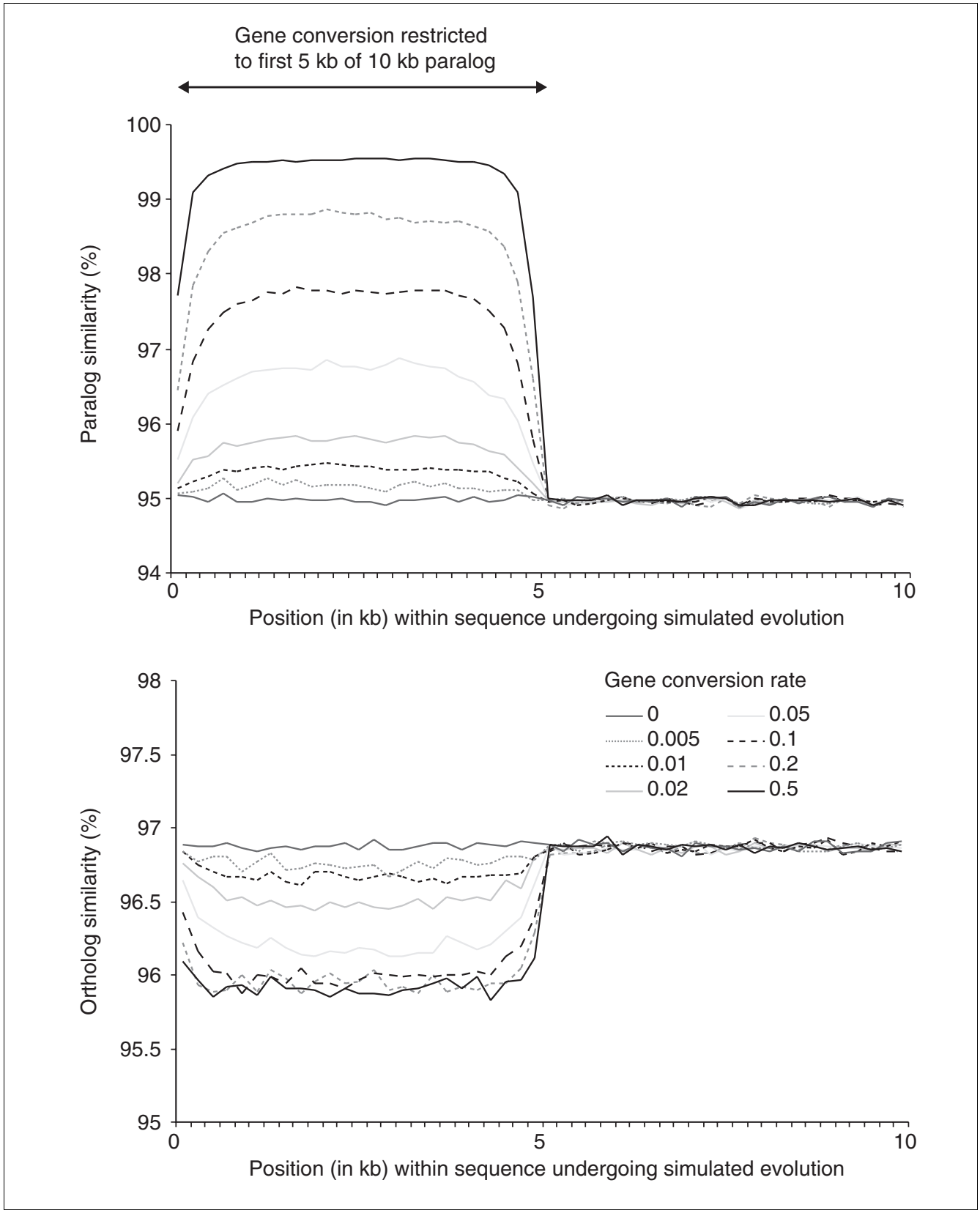

Figure 6 *(see legend on next page)*

Figure 6 *(see previous page)*
The effect of varying gene conversion rate on paralog and ortholog sequence similarity. Sequence similarities represent averages over 1,000 simulations. The two paralog pairwise comparisons (p1 vs d1 and p2 vs d2) are averaged together, as are the two ortholog comparisons (p1 vs p2 and d1 vs d2). Length of simulation 8×10^6 years; generation time 20 years; base substitution rate 4×10^{-8} per nucleotide per generation; gene conversion rate 0.1 (representing 4×10^{-5} per locus per generation - equivalent to 1.4×10^{-6} per site per generation); initial paralog sequence divergence 2%; gene conversion directionality 0.5 (that is, unbiased); mean gene conversion tract length 352 bp.

over time, and consequently, equilibrium-based methods [12] are inappropriate. These parameters are better estimated by studying gene conversion over shorter timescales, either by identifying individual gene conversion events [21] or by studying within-population variation [18,22]. Recent investigations of inter-paralog gene conversion processes in humans at the conversion hotspot flanking ID1 have estimated a gene conversion rate of $0.24 - 1.3 \times 10^{-3}$ events per generation and a mean tract length of 31 bp [18]. A recent study of inter-allelic gene conversion events at a recombination hotspot in sperm DNA has revealed higher rates of $5 - 16 \times 10^{-3}$ events per generation and longer mean tract lengths of 63-200 bp [21].

Using comparative sequence analysis to identify putative chromosomal rearrangement hotspots within segmental duplications

There are substantial technical difficulties associated with identifying NAHR hotspots directly from individual recombination events. However, should the association between an evolutionary history of gene conversion and recombination hotspot activity hold true across segmental duplications throughout the genome, then comparative sequence analysis seeking signatures of concerted evolution might allow the indirect identification of rearrangement hotspots. Most of the duplication events that generated the recombinogenic paralogous sequences in the human genome occurred during recent primate evolution [23], and so comparisons between human and mouse genomes are unlikely to be informative. In contrast, comparative sequences in great apes could facilitate the identification of putative rearrangement hotspots within segmental duplications through the application of the sliding-window analyses developed here. These analyses may reveal the presence of hotspots whose associated rearrangements have not been observed in the clinic as a result of their incompatibility with development to term. This may help to resolve why so few of the potentially recombinogenic segmental duplications are known to be associated with pathogenic consequences [24]. In this regard, it is unfortunate that the recently released draft chimpanzee genome sequence is unlikely to provide reliable comparative sequences of segmental duplications (M.E.H., unpublished work). This is partly due to the fact that the vast majority of paired reads in this whole-genome shotgun project come from short insert libraries and so do not provide sufficient positional information to assemble sequence reads to the correct copy of a segmental duplication. Further 'finished' sequences of chimpanzee chromosomes or targeted sequencing of specific segmental duplications will be required to conduct the *de novo* searches for rearrangement hotspots envisaged above.

Gene conversion is not the only process contributing to the homogenization of segmental duplications

The gene conversion processes described here are compatible with a model based on the mismatch repair of relatively short tracts of heteroduplex DNA formed during the processing of recombination intermediates. However, gene conversion as defined above is not the only process that can lead to homogenization of dispersed duplicated sequences. We previously identified homogenization events between these *AZFa*-HERVs that are incompatible with this model (Figure 10a), and suggested a double-crossover mechanism, which might either occur in a single meiosis, or via single unequal crossovers in different meioses - in other words, through the deletion of a duplicated intermediate (Figure 10b). The recent finding of these *AZFa* duplicated intermediates in the general population, and their apparent compatibility with fertility, suggest that the latter two-step scenario is more plausible [25]. This second mechanism for paralog homogenization has the potential to homogenize much longer tracts of segmental duplications than the mismatch repair-based mechanism. Thus it appears that at least two alternative mechanisms of paralog homogenization can operate at these *AZFa*-HERVs. The same may also be true of other segmental duplications underlying known genomic disorders where the reciprocal duplications of pathogenic deletions can be passed on from parent to offspring.

Consequences of inter-paralog gene conversion for evolutionary analyses and association studies

The demonstration that gene conversion can exaggerate levels of sequence divergence over and above those observed at non-duplicated loci in the same genome has important implications for evolutionary studies investigating variation in mutation rates between loci. Comparisons between sequence divergence at duplicated and non-duplicated loci are likely to be confounded by the presence of gene conversion in the former but not the latter. This factor may help to explain large discrepancies between recent estimates of the male-driven mutation rate in humans [26-30].

Against this background it is intriguing to note the recent study identifying reduced sequence divergence within long palindromic sequences undergoing gene conversion on the Y chromosome [12]. These palindromes have approximately 99.97% sequence similarity and our simulations suggest that

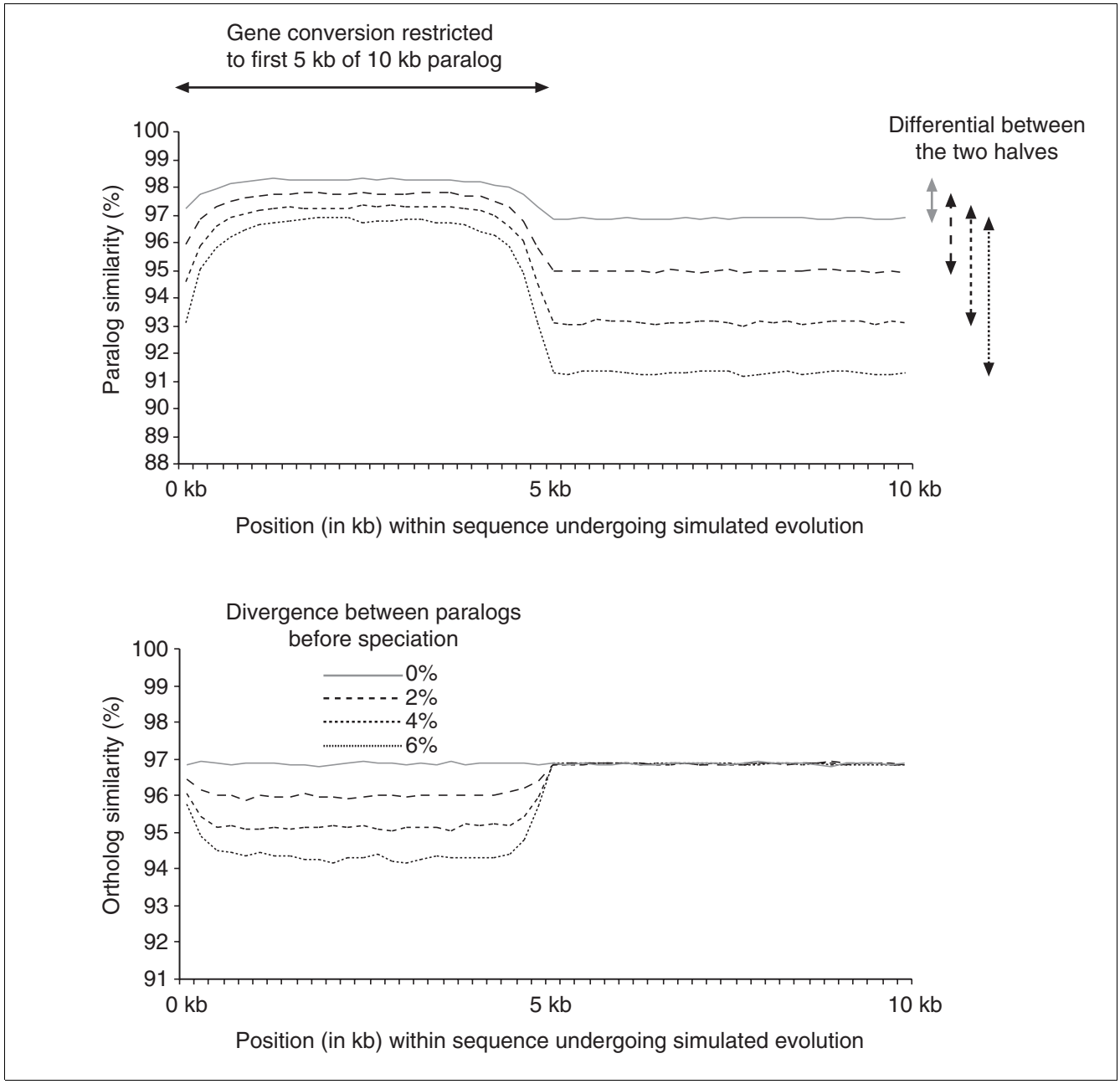

Figure 7
Paralog and ortholog sequence similarity under gene conversion, assuming different levels of sequence divergence between the paralogs before speciation. Sequence similarities represent averages over 1,000 simulations. The two pairwise paralog comparisons (p1 vs d1 and p2 vs d2) are averaged together, as are the two ortholog comparisons (p1 vs p2 and d1 vs d2). Length of simulation 8×10^6 years; generation time 20 years; base substitution rate 4×10^{-8} per nucleotide per generation; gene conversion rate 4×10^{-5} per locus per generation (equivalent to 1.4×10^{-6} per site per generation); mean gene conversion tract length 352 bp; gene conversion directionality 0.5 (unbiased).

as a consequence we should not expect gene conversion to elevate sequence divergence. The authors of this study proposed that by reducing sequence divergence, gene conversion might provide a means to maintain the functional integrity of genes residing in these palindromes. The simulations presented here indicate that gene conversion alone is unlikely to be capable of reducing orthologous sequence divergence. The reduction in sequence divergence is more likely to be due to an interplay of factors. For example, if gene conversion and base substitution were GC-biased in opposing directions, gene conversion events might preferentially maintain the ancestral state of a paralogous sequence variant.

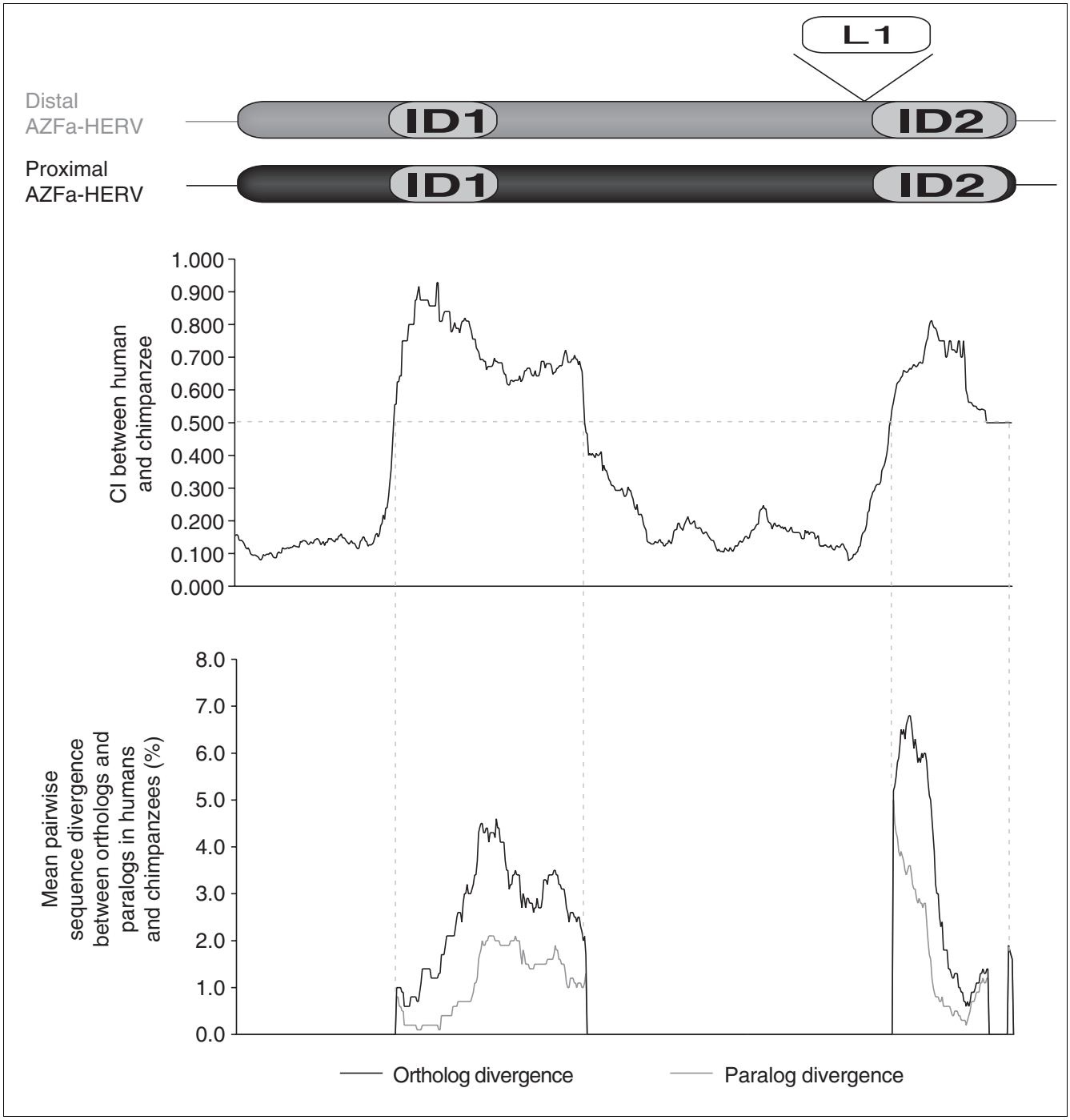

Figure 8
Sliding-window analyses across the alignment of human and chimpanzee *AZFa*-HERVs. Beneath the schematic alignment of the proximal and distal *AZFa*-HERVs (with the inserted L1 material excised) are sliding-window analyses showing how various sequence measures vary across the alignment. The measures applied are: the CI (see text) and the mean paralogous (averaged over both species) and orthologous (averaged over proximal and distal *AZFa*-HERVs) sequence divergences where the CI is greater than 0.5.

Gene conversion between paralogous sequences has important consequences for the analysis of human genomic diversity. It has recently been inferred from the preferential mapping of dbSNP entries to segmental duplications that some 100,000 dbSNP entries are not true single-nucleotide polymorphisms (SNPs), but are misassembled paralogous sequence variants (PSVs) [24]. However, this assertion does not take into account the possibility that gene conversion can

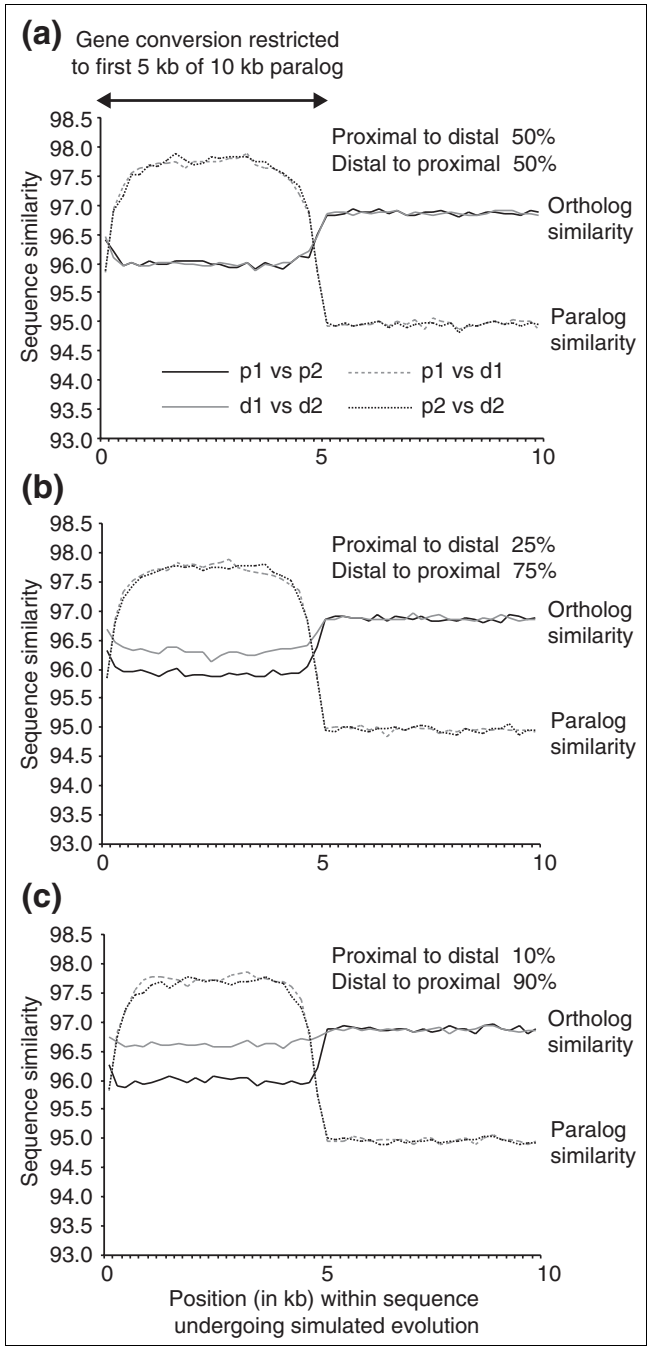

Figure 9

The effect of directionality of gene conversion on paralog and ortholog sequence similarity. Sequence similarities represent averages over 1,000 simulations. The effects of different ratios of gene conversion directionality are considered on the sequence similarity of paralogs p1 and d1, and p2 and d2, and of orthologs p1 and p2, and d1 and d2. **(a)** Conversion proximal to distal 50%, distal to proximal 50%; **(b)** proximal to distal 25%, distal to proximal 75%; **(c)** proximal to distal 10%, distal to proximal 90%. Length of simulation 8×10^6 years; generation time 20 years; base substitution rate 4×10^{-8} per nucleotide per generation; gene conversion rate 4×10^{-5} per locus per generation (equivalent to 1.4×10^{-6} per site per generation); mean conversion tract length 352 bp; initial paralog sequence divergence 2%.

elevate levels of sequence diversity at segmental duplications [31]. This suggestion draws support from both empirical [32] and theoretical [22,33,34] studies. Our analyses reveal that gene conversion provides a mechanism to elevate nucleotide variability by introducing new variants from paralogous sequences. Indeed, the simulations presented here for interspecific sequence divergence are formally no different from the processes operating at non-recombining loci to generate intraspecific diversity.

Gene conversion generates haplotypes with complex evolutionary histories. Further work is required to explore the haplotypic structure within segmental duplications. These results will have a bearing on the ability of SNP-based whole-genome association studies to identify etiologically important variants within segmental duplications, in which over 5% of known exons reside [24].

Materials and methods

Samples

Genomic DNAs from male chimpanzees and gorillas were kindly donated by Mark Jobling, and were supplemented by DNA extracted from chimpanzee blood supplied by the Institute of Zoology, UK.

Comparative sequencing

Each primate *AZFa*-HERV was amplified in its entirety, either in a single amplification using primers located in flanking single-copy sequence, or by using two overlapping long PCR reactions, each using one internal primer and one primer in the proximal- or distal-specific flanking sequences. All PCR amplifications were performed using either the DyNAzyme EXT (Finnzyme) or Expand 20 kb Plus (Roche) long PCR kits according to the manufacturers' instructions. Most of the sequence data was generated by shotgun sequencing these long PCR products (with more than 20× coverage) by generating paired reads from short insert libraries [35], that were subsequently assembled using PHRAP [36] and viewed using GAP4 [37]. For a minority of the data, a nested PCR strategy was used to generate shorter sequence templates from the proximal- and distal-specific *AZFa*-HERV long PCR products described above. These templates were fluorescently sequenced using primers at regular intervals designed from the human *AZFa*-HERV sequences, and the sequences assembled using either SeqEd (Applied Biosystems) or CAP (Contig Assembly Program [38]). Primer sequences are documented in Additional data file 1. The GenBank accession numbers for the sequences generated here are: AY573558-AY573561.

Two nomenclatures have been used to describe the potentially recombinogenic portions of the *AZFa*-HERVs [14,25]. The interval ID1 discussed here corresponds to segment A of Bosch and Jobling [25], and interval ID2 corresponds to the concatenation of segments B and C of Bosch and Jobling [25].

Segments B and C are separated by a single PSV, which appears not to be fixed in all human Y-chromosomal lineages (M.E.H., unpublished work).

Statistical analysis

Human sequences of *AZFa*-HERVs were extracted from finished BACs AC002992 and AC005820. Sequences were aligned using Se-Al [39]. The alignment is documented in Additional data file 2. Jukes-Cantor distances and neighbor-joining trees were calculated using Phylip [40]. Phylogenetic networks [41] were constructed using SplitsTree [42]. Sliding-window analyses of sequence similarities, concerted indices and directionality indices was performed using code written in Interactive Data Language 5.3 (Research Systems). These sliding windows were 500 bp long and were analysed at 15 bp intervals across the alignment.

Concerted evolution can be defined as the maintenance of homogeneity of nucleotide sequences among duplicated sequences within a species, although the nucleotide sequences change over time. We devised a statistic we call the concerted index (CI) to quantify this within-species similarity in relation to the observed variation between species. For a given alignment of paralogous and orthologous sequences in a pair of species, this statistic measures the ratio of the mean distance between orthologs (D_O) to the sum of the mean distance between orthologs and the mean distance between paralogs ($D_O + D_P$).

$$CI = D_O/(D_O + D_P)$$

If D_{p1p2} represents the distance between two orthologous sequences p1 and p2 in terms of the percentage of variant nucleotides between them, and D_{p1d1} represents the distance between two paralogous sequences p1 and d1 (using the sequence nomenclature of paralogs and orthologs shown in Figure 5), then the CI is calculated using the equation:

$$CI = ((D_{p1p2} + D_{d1d2})/2)/((D_{p1p2} + D_{d1d2})/2 + (D_{p1d1} + D_{p2d2})/2)$$

Consequently, when sequences are evolving in a concerted fashion, the mean distance between orthologs is relatively high, but the distance between paralogs is low, and the CI will tend to 1. This statistic is extended to the current situation where three species are represented by calculating the mean of the CI across each of the three possible pairwise comparisons. This equation could be extended to include any model of sequence evolution in the distance calculation. However, the high levels of similarity between the sequences being analysed here means that any such correction has negligible impact in this analysis.

The distribution of CI values is strongly bimodal in this analysis (data not shown), thus clearly distinguishing between those portions of the alignment that are undergoing concerted evolution and those that are not.

The DI measures the difference between orthologous sequence divergence at proximal and distal copies of a duplicated sequence, as a function of the mean orthologous sequence divergence.

$$DI = (D_{p1p2} - D_{d1d2})/((D_{p1p2} + D_{d1d2})/2)$$

Thus if there is strong proximal-to-distal directionality, the discrepancy between proximal and distal orthologous divergences will generate a more negative DI. High distal-to-proximal directionality will generate a more positive DI and with minimal directionality, the DI will tend towards 0.

Gene conversion simulations

Monte Carlo (stochastic) simulations were written in Interactive Data Language 5.3 (Research Systems) by M.E.H. (source code is available in Additional data file 3). The simulation models the post-speciation evolution of a pair of 10 kb duplicated sequences in two daughter species, for example human and chimpanzee. A model of sequence divergence is implemented in which each base in the four sequences is equally mutable, and is capable of undergoing reversions and parallel mutations. In addition, infrequent gene conversion events between paralogs are incorporated at random, but limited to the first half of the sequences. To represent the fact that the duplication event that generated the paralogs in the first place must have pre-dated the speciation event, the initial sequences of the paralogs are differentiated by a specified frequency of variant sites scattered at random (representing an initial sequence divergence of 0 to 6%), although orthologs in the two species are initially identical. This does not imply that gene conversion was absent before speciation, only that it is immaterial, given that variation between paralogs accumulates even in the presence of gene conversion, as homogenization is rarely perfect.

The gene conversion tract length is drawn at random from a geometric distribution [19], which is defined by a single parameter (ϕ). The probability (P) that the conversion tract length is n nucleotides long is given by the equation:

$$P(n) = (1-\phi)\phi^n$$

The mean tract length (T) can be simply related to ϕ by the equation:

$$T = \phi/(1-\phi)$$

For example, the value of ϕ determined by fitting a geometric distribution to the meiotic gene conversion tracts observed at the *rosy* locus in *Drosophila melanogaster* is 0.99717, which gives a mean tract length of 352 bp [19].

This conversion tract is positioned at random within the 5 kb portion of the duplicated sequence that is capable of undergoing gene conversion. The directionality of the gene conversion

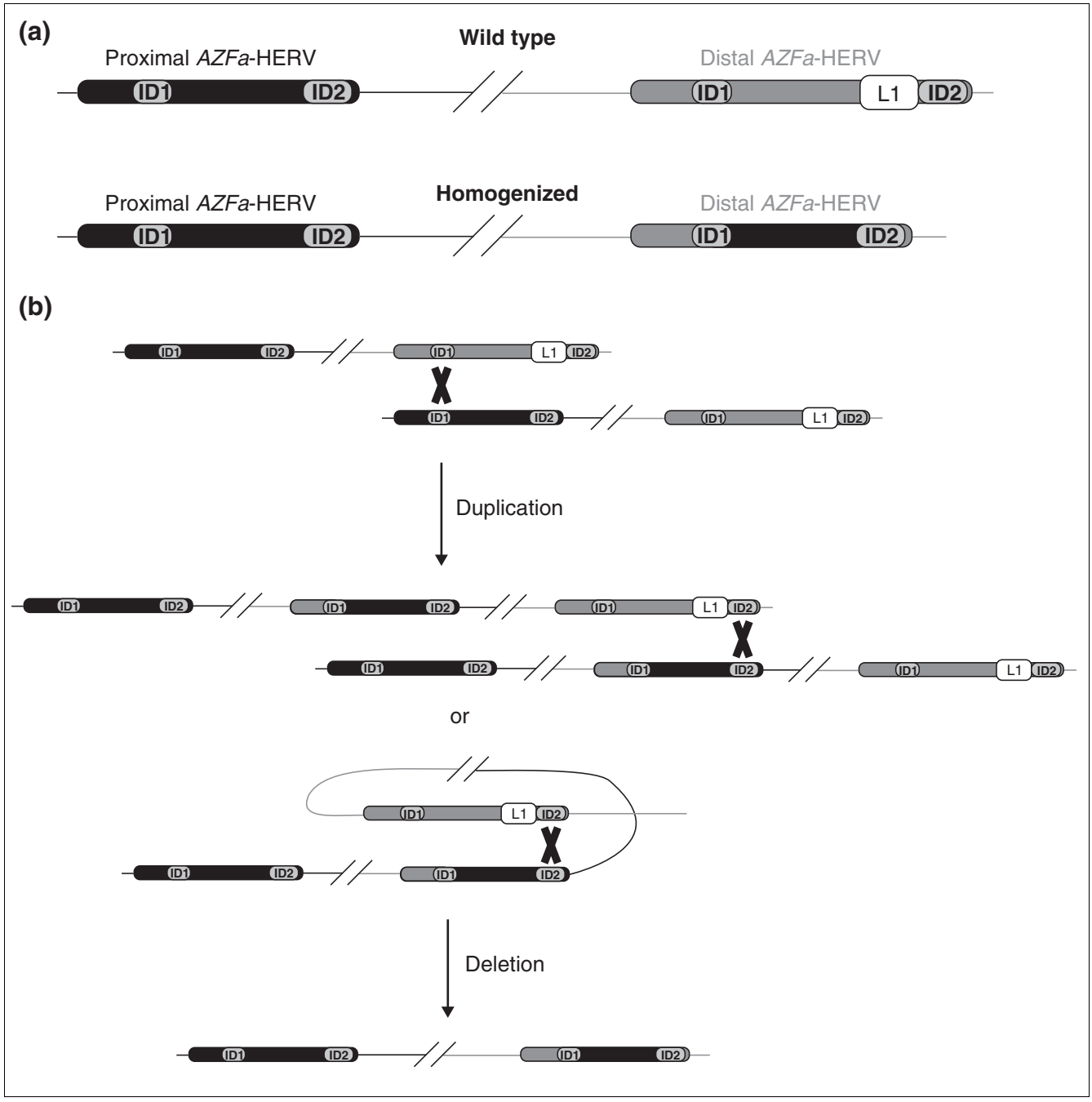

Figure 10
Structure and possible mechanism of formation of the homogenized *AZFa*-HERVs present in humans. **(a)** Comparison of the wild-type structure of *AZFa*-HERVs and the homogenized distal *AZFa*-HERV between ID1 and ID2 that has had 1.5 kb of L1 material removed and has been generated at least twice during recent human evolution [9]. **(b)** Possible two-step mechanism by which such a long portion of the distal *AZFa*-HERV could be homogenized via a duplicated intermediate. The constitutively haploid nature of the Y chromosome means that the substrates for the unequal crossing-over event that generates the duplicated intermediate must be sister chromatids. The unequal crossing-over event that results in deletion back to two copies might be either intra-chromosomal or between sister chromatids.

event is stochastically assigned according to a single parameter, x, which reflects the probability that the gene conversion donor is the proximal sequence, and the acceptor is the distal sequence. If x = 0.5, gene conversion occurs in either direction with equal probability.

The rate of gene conversion events (g) is defined relative to the base substitution rate per locus, such that if g = 0.1 then a gene conversion event between the two paralogs will occur on average after 10 base substitutions have arisen in each paralog. Therefore, given a base substitution rate of 4×10^{-8} per generation, a gene conversion rate of 0.1 is equivalent to a per locus rate of 4×10^{-5} per locus per generation (which represents a rate of 1.4×10^{-6} per site per sgeneration). After an amount of evolutionary time equivalent to a fixed number of generations, each simulation is halted and pairwise sequence similarities are calculated in non-overlapping 200 bp windows across the 10 kb sequence for each pair of orthologs and paralogs. Each simulation is replicated 1,000 times under the same parameters, and the sequence similarities for each pairwise comparison are averaged over all replications.

Additional data files

Additional data available with the online version of this paper comprise: additional data file 1, a table of primers used in amplifying and sequencing the *AZFa*-HERVs (Additional data file 1); additional data file 2, the alignment of proximal and distal *AZFa*-HERVs from human, chimpanzee and gorilla in FASTA format (Additional data file 2); and additional data file 3, the source code for simulations of gene conversion used in this study, written in Interactive Data Language (IDL)(Additional data file 3). See Materials and methods for details.

Acknowledgements

We thank Mark Jobling and Elena Bosch for discussions and sharing research materials, and are grateful to Mike Jackson for comments on an earlier manuscript. We are also grateful to Jane Rogers for her support. This work was supported by the McDonald Institute and the Wellcome Trust Sanger Institute.

References

1. Jeffreys AJ, Kauppi L, Neumann R: **Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex.** *Nat Genet* 2001, **29:**217-222.
2. Reiter LT, Murakami T, Koeuth T, Pentao L, Muzny DM, Gibbs RA, Lupski JR: **A recombination hotspot responsible for two inherited peripheral neuropathies is located near a mariner transposon-like element.** *Nat Genet* 1996, **12:**288-297.
3. Stankiewicz P, Lupski JR: **Genome architecture, rearrangements and genomic disorders.** *Trends Genet* 2002, **18:**74-82.
4. Jobling MA, Williams G, Schiebel K, Pandya A, McElreavey K, Salas L, Rappold GA, Affara NA, Tyler-Smith C: **A selective difference between human Y-chromosomal DNA haplotypes.** *Curr Biol* 1998, **8:**1391-1394.
5. Kauppi L, Sajantila A, Jeffreys AJ: **Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region.** *Hum Mol Genet* 2003, **12:**33-40.
6. Papadakis MN, Patrinos GP: **Contribution of gene conversion to the evolution of the human beta-like globin gene family.** *Hum Genet* 1999, **104:**117-125.
7. Waldman AS, Liskay RM: **Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology.** *Mol Cell Biol* 1988, **8:**5350-5357.
8. Elliott B, Richardson C, Winderbaum J, Nickoloff JA, Jasin M: **Gene conversion tracts from double-strand break repair in mammalian cells.** *Mol Cell Biol* 1998, **18:**93-101.
9. Blanco P, Shlumukova M, Sargent CA, Jobling MA, Affara N, Hurles ME: **Divergent outcomes of intrachromosomal recombination on the human Y chromosome: male infertility and recurrent polymorphism.** *J Med Genet* 2000, **37:**752-758.
10. Aradhya S, Bardaro T, Galgoczy P, Yamagata T, Esposito T, Patlan H, Ciccodicola A, Munnich A, Kenwrick S, Platzer M, *et al.*: **Multiple pathogenic and benign genomic rearrangements occur at a 35 kb duplication involving the NEMO and LAGE2 genes.** *Hum Mol Genet* 2001, **10:**2557-2567.
11. Hurles ME: **Gene conversion homogenizes the CMT1A paralogous repeats.** *BMC Genomics* 2001, **2:**11.
12. Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC: **Abundant gene conversion between arms of palindromes in human and ape Y chromosomes.** *Nature* 2003, **423:**873-876.
13. Zimmer EA, Martin SL, Beverley SM, Kan YW, Wilson AC: **Rapid duplication and loss of genes coding for the alpha chains of hemoglobin.** *Proc Natl Acad Sci USA* 1980, **77:**2158-2162.
14. Kamp C, Hirschmann P, Voss H, Huellen K, Vogt PH: **Two long homologous retroviral sequence blocks in proximal Yq11 cause AZFa microdeletions as a result of intrachromosomal recombination events.** *Hum Mol Genet* 2000, **9:**2563-2572.
15. Sun C, Skaletsky H, Rezen S, Gromoll J, Nieschlag E, Oates R, Page DC: **Deletion of *azoospermia factor a* (*AZFa*) region of human Y chromosome caused by recombination between HERV15 proviruses.** *Hum Mol Genet* 2000, **9:**2291-2296.
16. Shen PD, Wang F, Underhill PA, Franco C, Yang WH, Roxas A, Sung R, Lin AA, Hyman RW, Vollrath D, *et al.*: **Population genetic implications from sequence variation in four Y chromosome genes.** *Proc Natl Acad Sci USA* 2000, **97:**7354-7359.
17. Chen FC, Li WH: **Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees.** *Am J Hum Genet* 2001, **68:**444-456.
18. Bosch E, Hurles ME, Navarro A, Jobling MA: **Dynamics of a human inter-paralog gene conversion hotspot.** *Genome Res* 2004, **14:**835-844.
19. Hilliker AJ, Harauz G, Reaume AG, Gray M, Clark SH, Chovnick A: **Meiotic gene conversion tract length distribution within the *rosy* locus of *Drosophila melanogaster*.** *Genetics* 1994, **137:**1019-1026.
20. Taghian DG, Nickoloff JA: **Chromosomal double-strand breaks induce gene conversion at high frequency in mammalian cells.** *Mol Cell Biol* 1997, **17:**6386-6393.
21. Jeffreys AJ, May CA: **Intense and highly localized gene conversion activity in human meiotic crossover hot spots.** *Nat Genet* 2004, **36:**151-156.
22. Innan H: **A method for estimating the mutation, gene conversion and recombination parameters in small multigene families.** *Genetics* 2002, **161:**865-872.
23. Samonte RV, Eichler EE: **Segmental duplications and the evolution of the primate genome.** *Nat Rev Genet* 2002, **3:**65-72.
24. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE: **Recent segmental duplications in the human genome.** *Science* 2002, **297:**1003-1007.
25. Bosch E, Jobling MA: **Duplications of the *AZFa* region of the human Y chromosome are mediated by homologous recombination between HERVs and are compatible with male fertility.** *Hum Mol Genet* 2003, **12:**341-347.
26. Makova KD, Li WH: **Strong male-driven evolution of DNA sequences in humans and apes.** *Nature* 2002, **416:**624-626.
27. Bohossian HB, Skaletsky H, Page DC: **Unexpectedly similar rates of nucleotide substitution found in male and female hominids.** *Nature* 2000, **406:**622-625.
28. Ebersberger I, Metzler D, Schwarz C, Paabo S: **Genomewide comparison of DNA sequences between humans and chimpanzees.** *Am J Hum Genet* 2002, **70:**1490-1497.
29. Erlandsson R, Wilson JF, Paabo S: **Sex chromosomal transposable element accumulation and male-driven evolution.** *Mol Biol Evol* 2000, **17:**804-812.
30. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
31. Hurles ME: **Are 100,000 'SNPs' useless?** *Science* 2002, **298:**1509.
32. Giordano M, Marchetti C, Chiorboli E, Bona G, Momigliano Richiardi P: **Evidence for gene conversion in the generation of extensive polymorphism in the promoter of the growth hormone**

gene. *Hum Genet* 1997, **100:**249-255.
33. Innan H: **The coalescent and infinite-site model of a small multigene family.** *Genetics* 2003, **163:**803-810.
34. Nagylaki T: **Evolution of multigene families under interchromosomal gene conversion.** *Proc Natl Acad Sci USA* 1984, **81:**3796-3800.
35. McMurray AA, Sulston JE, Quail MA: **Short-insert libraries as a method of problem solving in genome sequencing.** *Genome Res* 1998, **8:**562-566.
36. Green P: **PHRAP assembly software.** [http://www.phrap.org].
37. Bonfield JK, Smith K, Staden R: **A new DNA sequence assembly program.** *Nucleic Acids Res* 1995, **23:**4992-4999.
38. Huang X: **A contig assembly program based on sensitive detection of fragment overlaps.** *Genomics* 1992, **14:**18-25.
39. Rambaut A: **Se-Al manual sequence alignment editor.** [http://evolve.zoo.ox.ac.uk/software.html?id=seal].
40. **PHYLIP** [http://evolution.gs.washington.edu/phylip.html]
41. Huson DH: **SplitsTree: analyzing and visualizing evolutionary data.** *Bioinformatics* 1998, **14:**68-73.
42. **SplitsTree: analyzing and visualizing evolutionary data** [http://www-ab.informatik.uni-tuebingen.de/software/splits/welcome_en.html]